REPORT DOCUMENTATION PAGE	Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
23-12-2010	Final Report	12-May-2009 - 11-Feb-2010

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Final Report: Meta-Learning Assistants Using a Novel Characterization of Data Landscapes	W911NF-09-1-0249
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 611102
6. AUTHORS	5d. PROJECT NUMBER
Ricardo Vilalta	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES	8. PERFORMING ORGANIZATION REPORT NUMBER
University of Houston Office of Grants & Contracts 316 E. Cullen Bldg. Houston, TX 77204 -2163	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	ARO
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) 56268-NS-II.1

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

Our project focused on the mathematical foundations needed to build meta-learning assistants. The overall goal is to know how we can acquire and exploit knowledge about learning (i.e., meta-knowledge) to understand and improve the performance of learning algorithms. To that end, our work focused on the following research problem: how can we decide if one single complex model, or rather a combination of simple models, is the best strategy to use when we face a supervised learning task? Our results show that a combination of simple models is often the

15. SUBJECT TERMS

machine learning, meta-learning, supervised learning, model selection, model complexity

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UU	UU	UU	UU		Ricardo Vilalta
					19b. TELEPHONE NUMBER 713-743-3614

Standard Form 298 (Rev 8/98)
Prescribed by ANSI Std. Z39.18

Report Title

Final Report: Meta-Learning Assistants Using a Novel Characterization of Data Landscapes

ABSTRACT

Our project focused on the mathematical foundations needed to build meta-learning assistants. The overall goal is to know how we can acquire and exploit knowledge about learning (i.e., meta-knowledge) to understand and improve the performance of learning algorithms. To that end, our work focused on the following research problem: how can we decide if one single complex model, or rather a combination of simple models, is the best strategy to use when we face a supervised learning task? Our results show that a combination of simple models is often the best choice, as a minimum increase in model complexity is equivalent to tenths of simple models.

List of papers submitted or published that acknowledge ARO support during this reporting period. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Number of Papers published in peer-reviewed journals: 0.00

(b) Papers published in non-peer-reviewed journals or in conference proceedings (N/A for none)

Number of Papers published in non peer-reviewed journals: 0.00

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts): 0

Peer-Reviewed Conference Proceeding publications (other than abstracts):

(2010) Vilalta R., Ocegueda-Hernandez F., Bagaria C.
A Conceptual Study of Model Selection in Classification: Multiple Local Models vs One Global Model. Second International Conference on Agents and Artificial Intelligence (ICAART-2010), Valencia, Spain.

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts): 1

(d) Manuscripts

Number of Manuscripts: 0.00

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED
Ricardo Vilalta	0.50
Chaitanya Bagaria	0.25
Francisco Ocegueda-Hernandez	0.25
FTE Equivalent:	**1.00**
Total Number:	**3**

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME
Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Final Report for Proposal No. 56268-NS-II
Meta-Learning Assistants Using a Novel Characterization of Data Landscapes

Ricardo Vilalta

Summary of Project Activities

Our project focused on the mathematical foundations needed to build meta-learning assistants. The overall goal is to know how we can acquire and exploit knowledge about learning (i.e., meta-knowledge) to understand and improve the performance of learning algorithms. To that end, our work focused on the following research problem: how can we decide if one single complex model, or rather a combination of simple models, is the best strategy to use when we face a supervised learning task? Our results show that a combination of simple models is often the best choice, as a minimum increase in model complexity is equivalent to tenths of simple models.

Figure 1 shows a diagram illustrating our main ideas. Traditional approaches to model selection vary complexity by jumping between model families F_i; every single model in the new family is able to create more flexible decision boundaries compared to any single model in the first family. Alternatively, complexity can vary by combining multiple models into a composite model (while fixing the complexity of each single model in the first family); every model in the new family F_{ik} is the result of combining k models from the first family F_i. New models are also more complex but due to the composite approach. The question is how do these two approaches compare? How much complexity is precisely increased with each approach? When combining k models, how far can k increase until complexity grows above the traditional approach of invoking single complex models? By answering these questions we open the possibility of including both approaches in the same model selection strategy, while expanding our understanding of learning-algorithm designs.

Our Theoretical Analysis

In what follows I provide a detailed description of our theoretical analysis (a full description can be found in our conference paper at ICAART (Vilalta et al., 2010)). We showed the conditions under which combining multiple local models is expected to be beneficial. In essence we wish to compare a composite model M_c to a basic global model M_b. M_c is the combination of multiple models. We assume M_b has VC-dimension h_b and M_c has VC-dimension h_c, which comes from the combination of k models, each of VC-dimension at most h, where we assume $h < h_b$.

The question we address is the following: how many models of VC-dimension at most h can M_c comprise to still improve on generalization accuracy over M_b, assuming both models have the same empirical error? The question refers to the maximum value of k that still gives an advantage of M_c over M_b. To proceed we look at the VC-dimension of h_c, which in essence is the VC-dimension of k-fold unions or intersections. It is an open problem to determine the VC-dimension of a family of k-fold unions (Reyzin, 2006; Blumer et al., 1989; Eisenstat and Angluin, 2007); recent work, however, shows that such a family of models has a lower bound of $\frac{8}{5}kh$, and an upper bound of $2kh \log_2 3k$ (it has been shown that $O(nk \log_2 k)$ is a tight bound (Eisenstat

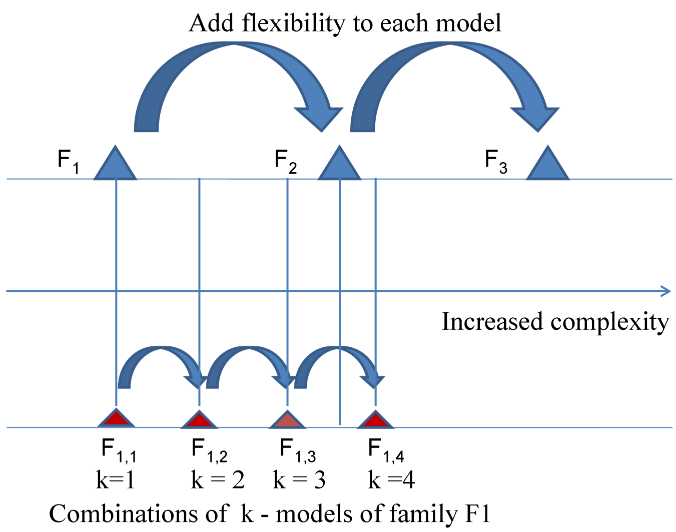

Figure 1: Two types of model selection. Top: Complexity is increased by looking at families of models F_i with increased flexibility in the decision boundaries. Bottom: Complexity is increased by combining k models while fixing the complexity of each model. F_{ik} stands for the combination of k models of family F_i. If we could compare both approaches –as in this example– we could say that model family F_{13} is less complex than family F_2, which in turn is less complex than family F_{14}.

and Angluin, 2007)). We begin our study with the lower optimistic bound, and assume the VC-dimension of h_c to be $\frac{8}{5}kh$. To solve the question above we equate Vapnik's guaranteed risk for both M_c and M_b:

$$\sqrt{\frac{\frac{8}{5}kh\left(\ln\frac{2N}{\frac{8}{5}kh}+1\right)-\ln(\frac{\eta}{4})}{N}} = \sqrt{\frac{h_b\left(\ln\frac{2N}{h_b}+1\right)-\ln(\frac{\eta}{4})}{N}} \quad (1)$$

where our goal is now simply to solve for k. After some algebraic manipulation we get the following:

$$c_1 k - k\ln k = c_2 \quad (2)$$

where c_1 and c_2 are constants:

$$c_1 = \ln 2N + 1 - \ln(\frac{8}{5}h) \quad (3)$$

$$c_2 = \frac{h_b}{\frac{8}{5}h}\left(\ln\frac{2N}{h_b}+1\right) \quad (4)$$

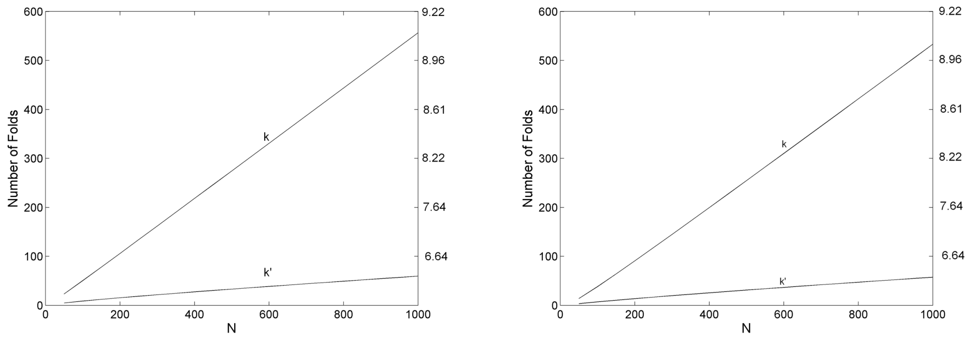

Figure 2: Left: A comparison of a compound model using k (k') support vector machines with polynomial kernels of degree one vs a simple support vector machine with a polynomial kernel of degree two; Right: same comparison except the simple support vector machine has a polynomial kernel of degree four. The degree of the polynomial kernel makes little difference in the results.

Equation 2 can be formulated as a transcendental algebraic equation. We can transform the equation as follows:

$$-c_2 k^{-1} e^{-c_2 k^{-1}} = -c_2 e^{-c_1} \tag{5}$$

To solve for k we can use Lambert's W function:

$$k = \frac{-c_2}{W(-c_2 e^{-c_1})} \tag{6}$$

where W can be solved using a numeric approximation.

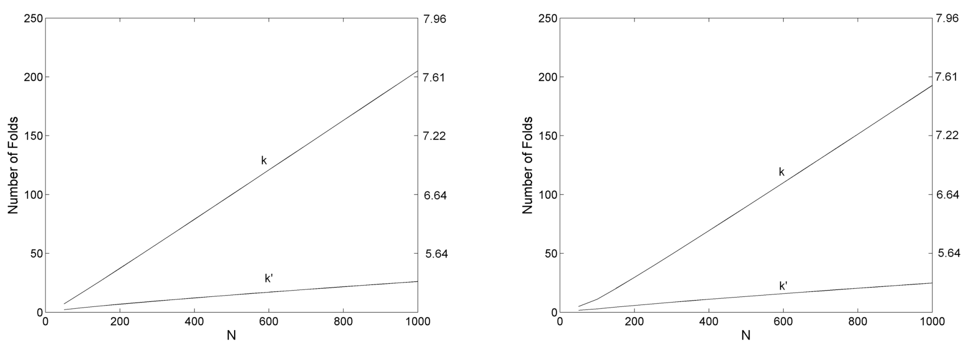

Figure 3: Left: A comparison of a compound model using k (k') support vector machines with polynomial kernels of degree two vs a simple support vector machine with a polynomial kernel of degree three; Right: same comparison except the simple support vector machine has a polynomial kernel of degree five. The degree of the polynomial kernel makes little difference in the results.

A similar analysis can be done using the upper bound of $h_c = 2k'h \log_2 3k'$, where we use k' to differentiate from the k used with the lower bound. After some algebraic manipulation we get the following equation:

$$c_3 \nu - \nu \ln \nu = c_4 \tag{7}$$

where $\nu = k' \ln 3k'$, and c_3 and c_4 are constants (only slightly different than before):

$$c_3 = \ln 2N + 1 - \ln(\frac{2h}{\ln 2}) \tag{8}$$

$$c_4 = \frac{h_b}{\frac{2h}{\ln 2}} \left(\ln \frac{2N}{h_b} + 1 \right) \tag{9}$$

Since equations 2 and 7 have the same form, ν has the same solution as k (equation 6):

$$\nu = \frac{-c_4}{W(-c_4 e^{-c_3})} = c_5 \tag{10}$$

We can then do the substitution back to k' to obtain the following:

$$k' \ln 3k' = c_5 \tag{11}$$

$$c_5 (k')^{-1} e^{c_5 (k')^{-1}} = 3c_5 \tag{12}$$

It is now possible to solve for k':

$$k' = \frac{c_5}{W(3c_5)} \tag{13}$$

To summarize, we have shown how to express the number of k-fold (and k'-fold) unions of models, each with VC-dimension h, such that the resulting compound model exhibits the same guaranteed risk as a single model with VC-dimension h_b (we assume of course that $h < h_b$). To clarify, we handle two bounds, k and k', because of our uncertainty in the VC-dimension of model unions. In principle we know there is a k'', that stands as the exact bound, below which M_c retains an advantage over M_b.

We can now study the effect on k (and k') as we vary parameters such as the size of the training set, or the VC-dimension of the models in the composite model M_c (as compared to the global model M_b). Figures 2 and 3 show plots on how the number of model unions varies with different values of N. In each case we take the compound model as the union of k (and k') support vector machines, where the simple global model is a single support vector machine. We assume the use of polynomial kernels where the VC-dimension of each model is defined as (Burges, 1998):

$$h = \begin{pmatrix} n + p - 1 \\ p \end{pmatrix} + 1 \tag{14}$$

where n is the dimensionality of the input space and p is the degree of the polynomial. In Figure 2 we assume a compound model with polynomial kernels of degree $p = 1$. The global model varies from a polynomial degree $p = 2$ (Figure 2-left) to a polynomial degree $p = 4$ (Figure 2-right). In all cases we assume $n = 5$. It is clearly observed that the value of k (k') increases linearly with N. As expected, k' corresponds to a less inclined line as the upper bound on the VC-dimension lowers the number of models we can place at the composite model while still generating less variance as the single model. In addition, a higher difference in VC-dimension (Figure 2-right) shows almost no difference in the shape of k (k') for different values of N. The right y-axis on each graph is the $\log_2$ of the values on the left y-axis; it is simply an indicator of how many local models we could arrange in a hierarchical structure (assuming a binary tree) while still generating less variance as the global model. We observe that for large values of N (e.g., $N > 500$), large hierarchies can be employed with little effect over the variance component.

Figure 3 assumes a compound model with polynomial kernels of degree $p = 2$. The single model varies from a polynomial degree $p = 3$ (Figure 3-left) to a polynomial degree $p = 5$ (Figure 3-right). The same effect is observed as before except under a different scale. In all graphs we observe a large advantage gained by the combination of many low-complex models as compared to a single model exhibiting higher complexity. The difference grows linearly on N and is considerable for $N > 500$.

Conclusions

Our study shows the advantage that comes when a piece-wise model fitting approach is used in classification. This is justified by the difference in the rate of complexity obtained by augmenting the number of boundaries per class (composite model) to the increase in complexity obtained by augmenting the capacity of a single global learning algorithm (classical approach). The former enables us to increase the model complexity in finer steps.

Our future goal is to use these results in building a model for the classification of classes and sub-classes in hierarchical learning problems. The key idea is to try to combine simple models as comprehensively as possible before any attempt is done to apply complex models for classification.

REFERENCES

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Eisenstat, D. and Angluin, D. (2007). The vc dimension of k-fold union. *Information Processing Letters*, 101(5):181–184.

Reyzin, L. (2006). Lower bounds on the vc dimension of unions of concept classes. *Technical Report, Yale University, Department of Computer Science*, YALEU/DCS/TR-1349.

Vilalta, R., Ocegueda-Hernandez, F., and Bagaria, C. (2010). A conceptual study of model selection in classification: Multiple local models vs one global model. In *Second International Conference on Agents and Artificial Intelligence (ICAART-2010).*